DOCUMENT RESUME

ED 171 725 TM 008 459

AUTHOR Cowie, Colin
TITLE Test Information. Using the Essay as an Assessment Technique. Set 77. Number One. Item 13.
INSTITUTION New Zealand Council for Educational Research, Wellington.
PUB DATE 77
NOTE 5p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; *Essay Tests; *Scoring; Testing; *Testing Problems; *Test Reliability; *Test Validity; *Writing Skills
IDENTIFIERS *Interrater Reliability

ABSTRACT

Certain testing procedures will overcome some of the problems associated with the use of essay tests. Essay tests may not validly indicate achievement because the questions included in the test may not fairly represent instructional content. Reliability may be a problem because of variations in examinee response in different situations, in test markers from one time to another, and between different markers. To offset validity problems, only those content areas which have been taught should be assessed. Writing ability should be tested at different times, using different types of essays and as many questions as possible. Questions should be phrased clearly and pretested before use. Intra-marker reliability is increased when graders score all examinee responses on a question, before moving to the next question. Inter-marker reliability is enhanced when the rating scale has seven to fifteen points, and clear distinctions are made between each mark. In analytic marking, values are assigned for each of the evaluation criteria--such as organization, style, content, and mechanics--and then totalled. In impressionistic or wholistic scoring, papers are sorted into three piles representing varying quality; then each pile is further sorted. (Six steps for judging reliability of scoring procedures are appended). (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Test Information

set 77
number one

Using the Essay as an Assessment Technique

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

K. C. Pickens

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Using the Essay as an Assessment Technique

By Colin Cowie

The main reason for the use of essays is probably the belief, usually unwarranted, that essay tests do a better overall job than objective tests of measuring educational achievement. The argument goes something like this. A major purpose of education is to prepare individuals to interact effectively with others in the realm of ideas; the basic tool of interaction is language; the student performs by speaking and writing; the essay is a sample of a student's performance, hence it provides a direct measure of educational achievement. The argument may be valid if the subject is English, but it is debatable when attainments in other subject areas are being considered.

Another reason for retaining essays is probably the belief that they have a more beneficial effect on student learning. If essay writing is never examined students will give it little attention. One of the major objectives of much of our teaching is surely the improvement of the skill of writing, and if essays contribute to this teaching goal then they serve a valuable purpose.

It may well be that essays are retained merely as a matter of tradition. Teachers have grown up with them and as far as they can tell essay examinations have done the job asked of them satisfactorily in the past.

Another reason advanced is that no effective alternatives are available. It is commonly thought that objective tests do not tap skills other than memory and recall. In fact, of course, objective tests which measure the higher order skills are possible. This is not to suggest that there are not certain skills which can only be tested by essay. The ability to generate one's own ideas, to structure a body of knowledge or to produce original thoughts are best assessed by essays.

Finally it may be that teachers believe that objective testing demands more technical skill than they possess. True, the construction of good objective test items is not a simple task, but it has been found that with guidance and practice, teachers can prepare tests of this type.

Given that essay questions will continue to be used by classroom teachers and external examiners, the purpose of this commentary is to delineate some of the problems relating to their use and to offer some suggestions for overcoming these.

Problems Relating to the Use of the Essay

Validity and Sampling

A measure of achievement is considered valid when it permits a sound, relevant assessment to be made of the extent to which a student has mastered a given set of instructional objectives. Very often students are assessed on how well they synthesise arguments, analyse situations, organise thoughts, develop original ideas, or express themselves clearly and correctly, even though their teachers have made little attempt to teach these skills.

It is known that students vary considerably in their writing ability from day to day and from topic to topic. Further, studies conducted by the Australian Council for Educational Research (ACER) have shown that students vary in their ability to write particular kinds of essays. Some may be strong on narrative and not so good at imaginative writing; good at exposition and not so good at descriptive writing.

Allied to the problem of validity is the problem of sampling. A major criticism of the essay examination has been that it samples a field of knowledge inadequately. Because extended written answers are required, only a small number of essay questions can be answered in a given time and thus a comprehensive and balanced coverage of the subject is not always possible. The teacher is faced with the dilemma of either increasing the adequacy of the sample by asking many different questions (mini-essays or paragraph answers) or asking few questions but those that probe deeply, whatever understanding the student has developed.

Then there is the problem of optional questions. It has often been argued that asking all students to write on the same question may penalise the student who has the misfortune to know little or nothing about the topic. This has led to the situation where a choice of questions is allowed, presumably to make the situation 'fairer'. However, if all questions involve essential elements of achievement why is it unfair to ask the student to answer all of them?

Reliability

If a test is reliable the score an individual obtains on that test is indicative of what he will do on other similar (but different) tests and further, the score is an indication of the individual's true level of accomplishment. The reliability of essay type tests or examinations can be affected by a number of factors.

(i) Examinee Reliability

A student's performance on a single test does not necessarily represent his 'typical' ability. There are factors operating within an individual, eg, mood, health, fatigue, etc, which might cause him to vary in his response from one test situation to another. Another factor contributing towards examinee variability is the matter of question interpretation. If a question is ambiguous then it is almost certainly unreliable. The examinee must have a clear understanding of the question and all examinee's must interpret the question the same way.

(ii) Intra-marker Reliability

If the same essay is marked by the same marker on two or more occasions the marks awarded will not necessarily be identical. Three separate time-related effects have also been shown to influence marks awarded by the same individual on separate occasions.

(a) The order in which essays are marked affects the marks awarded. That is, the mark an essay receives can be affected by the quality of the papers which precede it. A poor paper may receive a lower mark if it follows a good paper than if it were to follow other poor papers and vice-versa.

This aspect is aggravated when scoring schemes require that a specific distribution of marks are to be observed, eg if 10 essays are to be awarded the maximum mark, it may happen that the quota will be realised early. Consequently papers just as deserving of a top mark but which are marked last, are marked more stringently.

(b) Another version of the order effect can occur when an exam paper consists of a number of questions and the marker marks all questions before proceding to the next paper. The mark awarded to the first question may create in the marker's mind an expection in respect to the second, ie the 'halo effect' is said to operate.

(c) Then there is the temporal effect which has been observed in studies where large numbers of essays are marked over several days. The reliability of a marker drops significantly on

the last day of marking. This effect may not be so important in the classroom situation, though doubtless there are occasions when marking is spread over a number of days.

(iii) Inter-marker Reliability

It has frequently been found that when the same essays are marked by two different markers there are often considerable differences in the grades awarded. For example in a study carried out by the ETS in America 53 markers assessed 300 essays on a 9-point scale. It was found that 34% of the essays received all 9 possible grades and no essay was given fewer than 5 different grades out of 9. This inter-marker variability is due to a number of factors. Firstly, there is disagreement about what is to be assessed and how it should be measured. Markers also differ in their severity or leniency. One may tend to assign relatively high marks, another low marks. Markers also differ in the way they distribute marks throughout the scale —some avoid the extremes, while others use the complete range possible. Markers also differ in the relative values they assign to different papers.

While these kinds of differences in marking are clearly relevant to the marking of large-scale essay examinations they may be said to be of little importance when it comes to the marking of classroom tests. It can be argued that such tests are designed to measure a particular teacher's instructional objectives. The fact that other teachers in other settings might assign different scores to a paper is of no concern. While there is something to be said for this position there is also something to be said for the view that the assessment procedures should reflect a person's *true* achievement and not the idosyncrasies of the marker. And even in the school situation essay marks awarded by different markers are often compared.

Suggested Procedures for Overcoming Problem Areas

Validity and Sampling

(i) Unless you have an express purpose for assessing incidental learning, test only those things which have been taught.

(ii) If maximum validity is to be achieved it is obvious that we need several estimates of a student's writing ability. An ACER project directed by Jillian Maling Keepes (1) showed that for a true assessment of writing ability to be obtained, samples must be taken over time and include different kinds of essays.

(iii) Whenever the essay is utilized the number of questions asked should be increased so that more of the subject matter can be covered, and the questions should be structured so that the length of each answer is restricted, ie, more thought but less writing required.

(iv) Whether you offer a wide choice, a restricted choice, or no choice, will depend on what you are assessing and why. The teacher or examiner should keep in mind the notion that to be 'kind' to the student (by giving a wide choice) you are being 'unkind' to the marker (by making comparisons difficult). Unless all students are attempting the same questions it is very difficult to compare their efforts.

In summary, to improve the validity and sampling of the essay test, the teacher should:

(a) assess what has been taught and relate assessment to instructional objectives

(b) make assessments over time, topic and skill area

(c) increase the number of questions where possible and at the same time restrict the length of each answer

(d) require all students to answer all questions, ie, eliminate optional questions in essay testing.

Reliability

(i) Examinee Reliability

Little can be done to influence the extraneous factors operating within the individual. To obtain a 'true' measure of a student's writing ability a number of essays should be assessed. They should differ both in the kind of writing required and in length. This would help overcome the fact that performance varies very much from occasion to occasion and that writing ability varies significantly from topic to topic.

The easiest way to avoid question ambiguity it to impose sufficient structure on the question to guide the student in formulating his answer, while being careful not to confound what is being measured. Directions should be clear and precise; the form and wording of the question should state precisely what the student is expected to do. Pretesting the questions on a sample of students is one way to check the clarity of the question. Where this is not a practical solution, a reasonable alternative is to have the questions vetted by other teachers.

(ii) Intra-marker Reliability

Studies have shown that reliable results are obtained if the same question is marked over all papers before the marker proceeds to the next question. To alleviate the problem of the order in which questions are marked, the marker, after working through the set of papers, can change the order by shuffling them at random. The time effects are, of course, still present, but such a procedure does ensure that the bias is spread as evenly as possible and that no marks on any sub-set of papers are systematically depressed or inflated.

(iii) Inter-marker Reliability

Procedures for reducing inter-marker variability involve the use of a sufficiently fine scale, developing clear reference points and including multiple ratings.

Rather than using a five-point scale, greater reliability has been obtained if a scale having from 7 to 15 units is utilized: the greater the number of different points on a scale the more important it is to have clear notions about the characteristics of answers falling at each point. It is necessary to ensure that within certain limits all markers of a particular question use the scale in a similar way.

If the examination is concerned with knowledge, the points one would count as being relevant can be listed and anchored to set points in the scale. If the question is concerned with the student's ability to develop a synthesis of several ideas one may anchor the scale by developing model answers.

Another way of developing an anchor is to let the papers themselves determine the anchor by sorting out the papers according to a predetermined distribution that is to be used for every examination for the same class. Such a system has the advantage of ensuring there will be no shift in either the average score or in the spread of scores regardless of when the papers are marked. (Refer problem (ii) C).

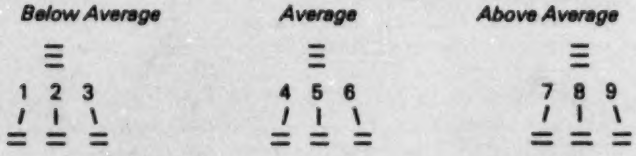
Teachers can never completely escape the systematic error reflecting their own special biases. The effects of variation in standard can be reduced if the procedures above are adopted. In addition personal bias can be reduced by working cooperatively with other teachers in the rating of essays. In general, two ratings, even if they are made rapidly will be preferable to a single rating. The sum of the two ratings will be more reliable than a single rating. In the Aorere College study, Elley et al (2) initially used 4 markers, each marking 4 essays for each pupil. They found there was little loss in reliability when this was reduced to 2 markers marking 3 essays, or 3 markers marking 2 essays. All the evidence suggests that multiple marking procedures result in a substantial rise in the consistency with which a student's work is assessed.

Scoring Schemes

A scoring scheme is simply a set of directions that governs the way in which marks are to be awarded to an essay. There are two general types of scoring schemes —(w)holistic or impressionistic marking and analytic marking.

1. *Impressionistic Marking*

(W)holistic marking requires that the essay be read through very quickly and that only a qualitative judgement of its merit be made This qualitative assessment is then converted to a mark or grade. As a first step in grading, gross judgements of the relative excellence of answers are made. Usually this involves sorting into three approximately equal piles. After the papers are reread subdivide each of these piles into three more piles, yielding nine piles in all. Then compare papers in piles 6 and 7 to ensure those in pile 7 are superior to those in pile 6. Do the same with piles 3 and 4. If more precision is required, and markers can make useful distinctions between papers an additional step could be carried out resulting in 9+ piles.



2. *Analytic Marking*

For analytic marking the marker is provided with a scale which stipulates those components of the essay that must be considered, and perhaps also, the maximum portion of the total mark which may be awarded to each component, eg, using a 16-point scale the following criteria are each to be alloted a maximum of 4 marks —content, organization, style, mechanics. A variation of this idea is to list the qualities which are to be considered in grading the essay together with a merit rating scale, for each feature.

Usually the ratings are given values of 1 to 5 and a composite mark for the essay is obtained by summing the checked values. The rationale for analytic marking is that, by directing markers' attentions to the same essay qualities and by regulating the distribution of marks, variability in marking may be reduced. Presumably the more explicit the scale the more reliable the results.

Surprisingly perhaps, research findings indicate that there are no significant differences in the reliability of marking between analytic and (w)holistic methods. That is, despite the relatively rapid rate with which essays may be marked (w)holistically, this scoring procedure has proved to be as reliable as any other. By contrast, analytic marking is very slow. Since it doesn't seem to matter which scoring scheme is used (when reliability is the basis for comparison) one should use the procedure which requires the least time. Analytic marking allows some diagnosis for further and future teaching emphasis.

If the essays are used solely for assessment purposes —as they would be when written for tests and examinations —the impressionistic marking is approximately twice as fast as analytic marking. If the essays are to be used for diagnostic and learning purposes then the time required for marking and annotating an essay by either scoring procedure becomes more equal.

How Reliable is Your Marking?

Follow the steps below to assess your own skill in rating essay tests. The procedure is adapted from Coffman (1972) (3).

Step 1. Have pupils identify themselves by number so that the ratings won't be influenced by the 'halo' effect.

Step 2. Rate the papers using your normal rating procedure. DO NOT record the ratings on the papers.

Step 3. Record the ratings on a separate grading sheet along with the corresponding identification numbers. Then put the papers away until there has been time to forget the details.

Step 4. Use the ratings to make whatever decisions are to be made and make a record of these decisions, eg, assign grades etc.

Step 5. Several weeks later, repeat the procedure again assigning ratings and applying the same decision rules.

Step 6. Count the number of pupils who are placed in a different decision group by the two procedures. What portion of the class is placed in a different group the second time? If the portion is small, the reliability is high; if the part is large, the reliability is low.

To illustrate, hypothetical data are used in the table below.

Table 1
Bivariate Frequency Distribution of Grades Based on Hypothetical First and Second Ratings of 25 Papers

First Rating	*Second Rating*				
	E	D	C	B	A
A				1	2
B			3	1	
C		3	7	1	1
D	1	1	1		
E	2	1			

$r_{12} = 0.77$

Note that 12 of the 25 students are in different grade categories on the two sets; 8 would have received lower grades and 4 higher grades. Notice, however, that only one would have received grades that differed by more than one category (a C on the first rating and an A on the second).

References

(1) Keepes, Jillian Maling & Rechter, Bernard: *English and its Assessment*. Melbourne, ACER, 1972

(2) Elley, W.B.; Barham, I.H.; Lamb, H.; & Wyllie, M.: 'The Role of Grammar in a Secondary School English Curriculum', *New Zealand Journal of Educational Studies*, 10, 26-42, 1975

(3) Coffman, W.E.: 'On the reliability of ratings of essay examinations', *NCME* Vol 3, No. 3, 1972

First issued in *set 77* number one. NZCER 1977.

Photograph by Jan Natua.